基于邻域粗糙集下知识划分的信息表降维 *

彭潇然 [a]，刘遵仁 [b]，纪 俊 [b]

(青岛大学 a. 数据科学与软件工程学院; b. 计算机科学技术学院, 山东 青岛 266071)

摘 要： Pawlak 粗糙集的知识约简包括对决策表的知识约简和对信息表的知识约简。作为 Pawlak 粗糙集的扩展，邻域粗糙集在针对决策表的属性约简方面应用广泛，而针对信息表的属性约简方面应用鲜少。为了设计一种适用于信息表的属性约简算法，根据 Pawlak 粗糙集的信息表知识约简标准，首先提出一种邻域粗糙集的信息表知识约简标准，然后根据这种标准，结合贪心思想，进一步提出了一种适用于聚类任务的信息表属性约简算法。与主成分分析(principal component analysis，PCA)算法相比，实验结果表明用该算法对数据集降维后，得到的属性约简集合的属性个数较多，K-means 算法根据属性集合进行聚类的精度较高。实验结果证明该算法能有效地应用于信息表的属性约简方面。

关键词： 降维；聚类；信息表；邻域粗糙集；属性约简

中图分类号： TP391　　doi: 10.3969/j.issn.1001-3695.2017.06.0650

Dimension reduction for information tables based on knowledge partition of neighborhood rough set

Peng Xiaoran[a], Liu Zunren[b], Ji Jun[b]

(*a. College of Data Science & Software Engineering*, *b. College of Computer Science & Technology*, *Qingdao University*, *Qingdao Shandong* 266071, *China*)

Abstract: Knowledge reduction of Pawlak rough set includes two parts: knowledge reduction for decision tables and knowledge reduction for information tables. As an extension of Pawlak rough set, neighborhood rough set is widely applied to attribute reduction for decision tables, but rarely applied to attribute reduction for information tables. In order to design an attribute reduction algorithm suitable for information tables, this paper first proposes a knowledge reduction criterion of neighborhood rough set for information tables based on the knowledge reduction criterion of Pawlak rough set. Then, according to this criterion, a new attribute reduction algorithm for information tables, applicable to clustering, is proposed with Greedy Strategy. Compared with Principal Component Analysis(PCA) algorithm, the experimental results show that by using this proposed algorithm to reduce dimensions of data sets, the number of attributes in the reduction sets is more, and the accuracy of K-means algorithm is higher according to the reduction sets, which proves this proposed algorithm can be effectively applied to attribute reduction for information tables.

Key Words: dimension-reduction; clustering; information table; neighborhood rough set; attribute reduction

0 引言

将性质、特征相似的样本归属到同一类别的过程被称为聚类。随着信息时代的高速发展，聚类所面临的难题不仅是“数据爆炸”问题，还有更重要的因高维数据产生的“维度灾难”现象[1]。因此，在尽可能保持知识表达能力不变的前提下，删除数据集中的冗余知识，在一定程度上消除噪声数据的干扰，对提高聚类算法的效率是十分有意义的。

粗糙集理论在海量高维复杂数据的预处理方面有着广泛的应用。经典的 Pawlak 粗糙集[2]通过等价划分对数据进行处理，但是这种等价划分只适用于离散型的数据，而在现实应用中，需要处理的数据往往是数值型的，这种局限滞缓了粗糙集理论的应用。为此，Zadeh[3]提出了信息粒化和粒度计算的概念。Lin[4]在信息粒化、粒度的基础上提出了邻域模型的概念。Hu 等人[5]基于邻域粒化和粗糙逼近的概念，进一步提出的邻域粗糙模型[6,7]可以处理数值型的数据，进一步拓展了粗糙集理论的应用范围。

知识约简是粗糙集理论的一个重要研究领域。Pawlak 粗糙集的知识约简包括对决策表和信息表的知识约简，决策表和信息表最大的区别在于决策属性的有无。作为 Pawlak 粗糙集的扩

基金项目：国家自然科学基金资助项目（61503208）

作者简介：彭潇然（1994-），男，湖北天门人，硕士研究生，主要研究方向为粗糙集理论（pxr1203@qq.com）；刘遵仁（1963-），男，副教授，博士，主要研究方向为粗糙集理论、数据挖掘、智能计算等；纪俊（1982-），男，副教授，博士，主要研究方向为数据挖掘、大数据技术、转化医学等.

展，邻域粗糙集在针对决策表的属性约简方面应用广泛[5,8-12]。例如，Hu[5]提出了基于前向贪心的决策表属性约简算法；刘[8]结合粒子群算法提出了高维数据集快速约简算法；Liu[9]根据映射划分提出了快速决策表属性约简算法等等。这类算法依赖于邻域粗糙集的正域计算，而正域计算依据决策属性，即样本类别是已知的情况下，所以针对决策表的属性约简算法并不适用于信息表。

目前，邻域粗糙集针对信息表的属性约简方面应用鲜少。为了扩展邻域粗糙集对信息表的应用，基于邻域粗糙集下的知识划分，设计一种针对信息表，适用于聚类任务的属性约简算法，本文根据 Pawlak 粗糙集对信息表进行知识约简的标准，首先提出了一种适用于数值型信息表的知识约简标准，然后在此标准的基础上，结合贪心思想，进一步提出了一种基于前向贪心的信息表属性约简算法(fast attribute reduction algorithm for information table, FARAIT)，最后通过与为无监督性学习的主成分分析(PCA)算法相比，实验验证了本文算法的有效性。

1　相关概念

1.1　Pawlak 粗糙集下的知识划分[2]

经典的 Pawlak 粗糙集认为知识是有粒度的，它是一种对论域中样本进行分类的能力。

定义 1　信息粒。设 $U=\{x_1,x_2,\ldots,x_n\}$ 为样本的非空有限集合，称为论域。论域 U 的任何一个子集 $X\subseteq U$，称为 U 的一个概念或范畴，且每一个概念表示 U 的一个信息粒。

定义 2　不可分辨关系。给定一个论域 U 和 U 上的一簇等价关系 S，若 $P\subseteq S$，且 $P\neq\varnothing$，则 $\bigcap P$（P 中的所有等价关系的交集)仍然是论域 U 上的一个等价关系，且称为 P 上的不可分辨关系，记为 $\mathrm{IND}(P)$，且 $\forall x\in U$，$[x]_{IND(P)}=[x]_p=\bigcap_{\forall R\in P}[x]_R$。

根据以上定义，对于知识库 $K=(U,B)$，B 是 U 上的一簇等价关系，若 $E\subseteq B$，且 $E\neq\varnothing$，$U/\mathrm{IND}(E)=\{[x]_{\mathrm{IND}(E)}\mid\forall x\in U\}$ 表示与等价关系 $\mathrm{IND}(E)$ 相关的知识，即论域 U 根据等价关系 E 被划分成了若干个等价类（信息粒）。例如，$U_1=\{x_1,x_2,x_3,x_4,x_5\}$ 根据关系 E_1 被划分为 $\{\{x_1,x_2\},\{x_3\},\{x_4,x_5\}\}$。

1.2　邻域粗糙集下的知识划分[5]

Pawlak 粗糙集通过等价关系保证了粒度计算的进行，这种等价关系在离散型的数据集中可以直接构造，在数值型的数据集中却不能。作为 Pawlak 粗糙集的扩展，邻域粗糙集在处理数值型数据集时得到了很好的应用。

定义 3　度量计算。给定 n 维实数空间 R^n，对于空间中的任意两个点 $x_i=(x_{i1},x_{i2},\ldots,x_{in})$ 和 $x_j=(x_{j1},x_{j2},\ldots,x_{jn})$，定义 $d(x_i,x_j)$ 是 R^n 上的一个度量计算，满足：

$$d(x_i,x_j)=\left(\sum_{p=1}^{n}\left|x_{ip}-x_{jp}\right|^2\right)^{\frac{1}{2}}$$

定义 4　邻域粒化。在实数空间上，定义样本的非空有限集合 $U=\{x_1,x_2\ldots,x_n\}$，且称 U 为论域。定义 U 上的样本 x_i 的 δ-邻域为 $\delta(x_i)=\{x_j\mid x_j\in U,d(x_i,x_j)\le\delta\}$，其中 $\delta\ge 0$。$\delta(x_i)$ 称作由 x_i 生成的 δ-邻域信息粒子，简称为 x_i 的邻域粒子。

根据以上定义，对于知识库 $K=(U,B)$，若 $E\subseteq B$，且 $E\neq\varnothing$，则论域 U 根据关系 E 被划分成了若干个邻域信息粒子。例如，$U_2=\{x_1,x_2,x_3,x_4,x_5\}$ 根据关系 E_2 被划分为 $\{\{x_1,x_2\},\{x_2,x_1,x_3\},\{x_3,x_2\},\{x_4,x_5\},\{x_5,x_4\}\}$。

1.3　粗糙集的知识表达系统[2]

定义 5　知识表达系统。称四元组 $\mathrm{KRS}=(U,A,V,f)$ 是一个知识表达系统。其中 $U=\{x_1,x_2,\ldots,x_n\}$ 为样本的非空有限集合，称为论域；$A=\{a\mid a\in A\}$ 为属性的非空有限集合；$V=\bigcup_{a\in A}V_a$ 表示全体属性的值域，V_a 为属性 $a\in A$ 的值域；f 表示 $U\times A\to V$ 的一个映射，称为信息函数。

在知识表达系统 KRS 中，令 $A=C\cup D(C\cap D\neq\varnothing)$，其中 C 称为条件属性，D 称为决策属性。若 $D\neq\varnothing$，则知识表达系统称为决策表(decision table,DT)，否则称为信息表(information table,IT)。一般来说，决策表用于分类任务，信息表用于聚类任务。

2　本文提出的信息表知识约简标准

在保持知识表达系统的知识表达能力不变的前提下，删除知识系统中的冗余知识，称为知识约简。对于一个信息表而言，不同的知识就是条件属性的不同集合，冗余知识就是可删除的属性。

定义 6[2]　Pawlak 粗糙集的信息表知识约简标准。给定一个信息表 $\mathrm{IS}=(U,C,V,f)$，$\forall B\subseteq C$，$\forall a\in B$，若论域 U 根据属性集合 $B-\{a\}$ 的划分和根据属性集合 B 的划分不一致，则称属性 a 是属性集合 B 中不可删除的属性，反之则称属性 a 是属性集合 B 中可删除的属性。

根据定义 2 和 4 可知，邻域粗糙集下的知识划分和 Pawlak 粗糙集下的知识划分不同，相较 Pawlak 粗糙集，邻域粗糙集下的知识划分将论域上的等价关系变成了覆盖关系。如果对于邻域粗糙集下的知识划分直接引用定义 6，由于论域中的每个样本都形成了一个邻域信息粒子，且每个邻域信息粒子中样本不一，则在判断定义 6 中两种划分是否一致时，比较过程不便且计算量较大。针对这个问题，考虑到邻域粗糙集下知识划分的特性，本文将知识划分的变化做一个等价转换，提出定理 1。

定理 1　对于一个信息表而言，对于某个 δ 取值，在邻域粗糙集知识约简的过程中，判断论域的知识划分是否一致可以转换为判断互为邻域样本的对数是否一致。

证明　对于一个如表 1 所示的数值化信息表，根据全属性集合 C，可以将论域 $U=\{x_1,x_2,\ldots,x_n\}$ 映射成实数空间上的 n 个 m 维向量的集合，即 $x_n=(\mathrm{value}_{n1},\mathrm{value}_{n2},\ldots,\mathrm{value}_{nm})$。同理，根据部分属性集合 B $(\forall B\subseteq C)$，亦可以将其映射成 n 个 $t(1\le t\le m)$ 维向量的集合。

表 1　一个数值化信息表

样本	属性 1	属性 2	...	属性 m
x_1	value_{11}	value_{12}	...	value_{1m}
x_2	value_{21}	value_{22}	...	value_{2m}
...	...	...	...	...
x_n	value_{n1}	value_{n2}	...	value_{nm}

对于根据属性集合 B 得到的 n 个向量集合，在实数空间上根据邻域粒化的概念，论域 U 能得到一个对应的知识划分，即一个个邻域信息粒子，根据属性集合 $B-\{a\}$ $(\forall a\in B)$ (各向量同时减少 a 对应的维度)同样可以得到一个新的知识划分，判断属性 a 是否可删除的关键在于判断两种知识划分是否一致。可知，知识约简在实数空间上是一个降维的过程。

为了方便说明，用 5 个向量代表 n 个 t 维向量，即 $U=\{x_1,x_2,x_3,x_4,x_5\}$，用二维空间代表 t 维的高维空间，用一维空间代表降维后的 $t-1$ 维的低维空间，以此对约简时知识划分的变化过程进行分析。其中，减少的维数对应待判断是否可删除的属性。

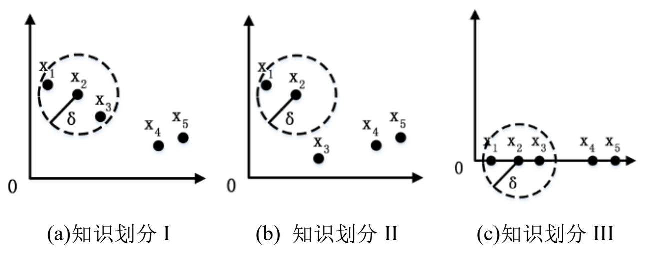


图 1　U 在降维时的知识划分过程

如图 1 所示，虚线圆圈代表邻域，(a)、(b)表示高维空间，(c)表示低维空间。在如(a)所示的高维空间中，U 被划分为 $\{\{x_1,x_2\},\{x_2,x_1,x_3\},\{x_3,x_2\},\{x_4,x_5\},\{x_5,x_4\}\}$，互为邻域样本的对数为 3，分别是 $\{(x_1,x_2),(x_2,x_3),(x_4,x_5)\}$。然后，不考虑待判断是否可删除属性对应的维数，得到如(c)所示的低维空间和相应的 U 的划分，比较(a)、(c)，发现前后划分是一致的，互为邻域样本的对数是相同的，这说明该属性是可删除的；若高维空间是如(b)所示的高维空间，U 被划分为 $\{\{x_1,x_2\},\{x_2,x_1\},\{x_3\},\{x_4,x_5\},\{x_5,x_4\}\}$，互为邻域样本的对数为 2，分别是 $\{(x_1,x_2),(x_4,x_5)\}$。此时比较(b)、(c)，发现前后划分是不一致的，相比前者，互为邻域样本的对数增加了，与前者不相同，这说明该属性是不可删除的。

以样本 x_2 为代表分析降维过程：在降维过程中，原本在高维空间中属于 x_2 的圆圈(邻域)内的样本点，因为减少了一维，其与 x_2 的欧式距离进一步减少，所以在降维后的低维空间中肯定仍在圆圈中；原本不在圆圈内的样本点，因为其与 x_2 的欧式距离的减少，在降维后会出现仍不在圆圈内和在圆圈内这两种情况。即对于互为邻域样本的对数而言，只有保持不变和增加这两种情况这两种情况，且分别对应属性可删除和不可删除这两种结果。这说明，对于某个 δ 取值，在邻域粗糙集知识约简的过程中，判断论域的知识划分是否一致可以转换为判断互为邻域样本的对数是否一致。

相对而言，计算互为邻域样本的对数是较容易的，利用上三角矩阵或下三角矩阵就能完成该计算。根据定理 1，本文能得到一种较直观且方便的判断知识划分是否变化的依据，据此提出一种邻域粗糙集的信息表知识约简标准。

标准 1　邻域粗糙集的信息表约简标准。给定信息表 $\text{IS}=(U,C,V,f)$，$\forall B\subseteq C$，$\forall a\in B$，基于邻域粗糙集，设在论域 U 根据属性集合 B 划分的邻域信息粒子中，互为邻域样本的对数为 N_{bef}，在论域 U 根据属性集合 $B-\{a\}$ 划分的邻域信息粒子中，互为邻域样本的对数为 N_{aft}。若 $N_{aft}==N_{bef}$，即前后划分一致，则可判定属性 a 是属性集合 C 中可删除的属性。若 $N_{aft}>N_{bef}$，即前后划分不一致，则可判定属性 a 是属性集合 C 中不可删除的属性。

3　基于前向贪心的信息表属性约简算法

定义 7[1]　属性约简。给定信息表 $\text{IS}=(U,C,V,f)$，$B\subseteq C$，若论域 U 根据属性集合 B 的划分和根据属性集合 C 的划分一致，且属性集合 B 中任意属性 a 均是不可删除属性，则称属性集合 B 是 C 的一个属性约简。

根据定义 7 可知，针对信息表的属性约简算法的目的是找到一个部分属性集，论域根据其形成的知识划分与根据原属性集形成的知识划分相同。

根据定义 7 和标准 1，“盲删法”是一种较易想到的算法。其思想是：对于一个数据集的原属性集，挑选其中一个属性作是否可删除判断，若可以，则从属性集中删除属性，对并新属性集重复挑选且判断的动作；若不可以，则对属性集中剩下的属性作判断。终止的条件是属性集中所有元素均是不可删除属性。

分析上述“盲删法”的时间复杂度。设在该算法下，假设某一数据集有 U 个样本， m 个属性，依次对各属性进行判断，约简结果中包含 k 个属性。则算法最好的情况是前 m-k 个属性为可删除属性，最坏的情况是后 m-k 个属性为可删除属性，两种时间复杂度的式子可简要表示为：

最好情况下：

$$1\cdot(m-1)|U|+1\cdot(m-2)|U|+\ldots+k\cdot(k-1)|U|$$

最坏情况下：

$$m\cdot(m-1)|U|+(m-1)\cdot(m-2)|U|\ldots+k\cdot(k-1)|U|$$

即对含有 n 个属性的属性集而言，对其中的某个属性进行可否删除判断时，需要在 n-1 维的实数空间上对 U 个样本进行知识划分的计算。其中，最好情况是每次删除时只需一次判断，最坏情况是每次删除时需要 n 次判断。

“盲删法”是一个降维的过程，即实数空间上从 m 维至 k 维的过程，由于初始维数很高，且对一个数据集而言，k 相较 m 而言通常是一个很小的值，当 m 较大且 k 较小时，整个过程的计算量是很大的。如果将降维过程改进为升维过程，即实数空间上从 0 维至 k 维的过程，则计算量会缩减很多。

根据以上思路，借鉴文献[5]中一种前向贪心的算法思想，根据第 2 节中的标准 1，提出一种前向贪心的信息表属性约简算法(Fast attribute reduction algorithm for information table, FARAIT)，用于对信息表求得一个最优或次优的属性约简。FARAIT 算法提出一种属性重要度概念，对于一个信息表 $\mathrm{IS}=(U,C,V,f)$ 而言，初始化约简集合为空集，每次贪心选取重要度最大的待选属性加入约简集合中。

定义 8　属性重要度。给定信息表 $\mathrm{IS}=(U,C,V,f)$，$B\subseteq C$，且 $B\neq\varnothing$，$\forall a\in C-B$，定义 a 相对于 B 的重要度为:

$$SIG(a,B)=\left[Card(B)-Card(B\cup a)\right]/Card(B)$$

其中，$Card(B)$ 表示根据属性集 B 形成的邻域信息粒子中互为邻域样本的对数。

在约简集合中加入属性时，实数空间上的升维过程会让各样本间的距离变大，使互为邻域样本的对数变少，用减少量衡量各属性的重要度。属性重要度可以理解为通过该属性区分样本的能力。例如，一个信息表按照原属性集能聚类成“男生”和“女生”两个类别，在贪心选择时，“身高”和“脸型”这两个属性的重要度不同。在实数空间上 “脸型”数值的分布较均匀，“身高”数值的分布呈两边逐步向中间聚集。相比之下，在约简集合中加入“身高”属性能使互为邻域样本的对数减少得更多，本文认为其区分能力更强，属性重要度更高。

相比“盲删法”，前向搜索算法能够确保重要的属性首先被加入到约简中，从而不损失重要的属性，而“盲删法”却难以保证这个结果。因为对于有大量冗余属性的信息表而言，即使那些重要的属性被删除也不一定会降低整个系统的区分能力，因此，系统最终可能保留了大量区分能力很弱、但作为一个整体依然能够保持原始数据的分辨能力的属性，而不是少量区分能力很强的属性。

FARAIT 算法的具体策略如下：初始化属性约简集合为空集，每次对不属于属性约简集合中的属性进行重要度计算，选取重要度值最大的属性加入约简集合中，直到所有剩余属性的重要度为 0，此时，根据约简集合形成的邻域信息粒子中互为邻域样本的对数不再变化，本文认为对应的知识划分与在原属性集下形成的知识约简一致或相似，任意剩余属性在标准 1 下均是可删除属性。如算法 1 所示。

算法 1

Input: $\mathrm{IS}=(U,C,V,f)$.

Output: 属性约简 red .

1: 初始化 $red=\varnothing$

2: if $Card(red)=0$ //此处定义 $\mathrm{Card}(\varnothing)=|U|^2/2$

  go to 6;

end

3: 对任意 $a_i\in C-red$ ，计算:

  $SIG(a_i,red)=[Card(red)-Card(red\cup a_i)]/Card(red)$；

4: 选择 a_k ，其满足:

  $SIG(a_k,red)=\max_i(SIG(a_i,red))$；

5: if $SIG(a_k,red)>0$

  $red\leftarrow red\cup a_k$；

  go to Step 2;

else

  go to Step 6;

end

6: return red；

设在该算法下，假设某一数据集有 U 个样本， m 个属性，依次对各属性进行判断，约简结果中包含 k 个属性。则 FARAIT 算法时间复杂度的式子可简要表示为

$$m\cdot 1\cdot|U|+(m-1)\cdot 2\cdot|U|+\ldots+(m-k)(k+1)|U|$$

可见，因为引用了贪心思想，FARAIT 算法能用较短的时间得到一个最优或次优的属性约简。

4　实验分析

作为一种有效的无监督性降维算法， PCA 算法在人工智能、模式识别、图像处理等方面得到了广泛的应用[13-15]。实验首先讨论 δ 的取值对 FARAIT 算法的的影响，确定较为合适的 δ 的取值；然后，在此 δ 的取值下，用 FARAIT 算法和 PCA 算法分别对数据集进行降维处理，用 K-means 算法对降维后的数据集进行聚类；最后，实验将比较得到的属性约简个数和 K-means 算法的聚类精度。其中，K-means 算法用于检验两种算法的降维效果，且 K-means 算法的初始聚类中心个数设置为数据集提供的类别数。

4.1　实验环境

UCI(University of California Irvine)(http://archive.ics.uci.edu/ml/)提供了一系列用于测试的标准数据集。本文从 UCI 数据集中挑选了 7 个数值型数据集，其中，每个数据集提供了条件属性和决策属性。

表 2　数据集描述

数据集	样本数	属性数	类别数
wine	178	13	3
WDBC	569	30	2
sonar	208	60	2
ionosphere	351	33	2
credit approval	690	13	2
german credit	1000	24	2
WPBC	198	33	2

本次实验在一台 Intel(R) Core(TM) i5 CPU 和 4GB 内存的 PC 机上，采用 Windows 7 环境下的 MATLAB R2016b 进行算法仿真。

4.2　FARAIT 算法的实验分析

本部分将对 FARAIT 算法进行具体分析。在不考虑数据集中决策属性的前提下，首先在不同的 δ 取值下用 FARAIT 算法对数据集 wine、WDBC、sonar 进行降维，然后用 K-means 算

法对降维后的数据集和原数据集分别进行聚类。记录属性约简的个数，且将得到的两种聚类结果和数据集中提供的决策属性进行对比，统计各自正确聚类的样本个数并计算正确率(精度)，从而确定较为合适的 δ 的取值。

4.2.1 δ 的取值和属性约简个数

根据定义 4 可知，δ 的取值直接影响着属性约简的结果。在不同的 δ 取值下，算法得到的属性约简不同，这会造成根据属性约简进行聚类后，所得的聚类精度不同。本文在区间 $[0,1]$ 上，按 0.02 增进，共取得 51 个 δ 取值，记录在不同的 δ 取值下对应的属性约简个数和聚类精度，其中 K-means 算法执行 20 次，聚类精度的最后结果取均值。

δ 的取值与属性约简个数的关系如图 2 所示。

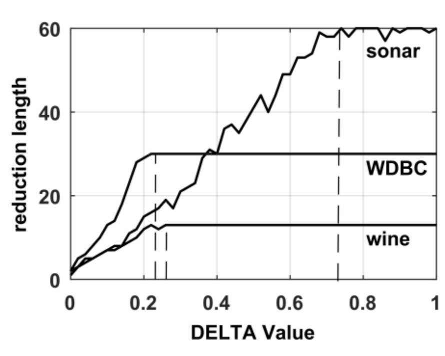


图 2 δ 的取值与属性约简个数的关系

如图 2 所示，对同一数据集，δ 的取值不同时，所得的属性约简个数也不同。从 0 开始，随着 δ 取值的增大，所得属性约简的个数增大，直到增大到和原属性集长度一致时稳定，称此时的 δ 取值为饱和点。

4.2.2 δ 的取值和 K-means 算法的聚类精度

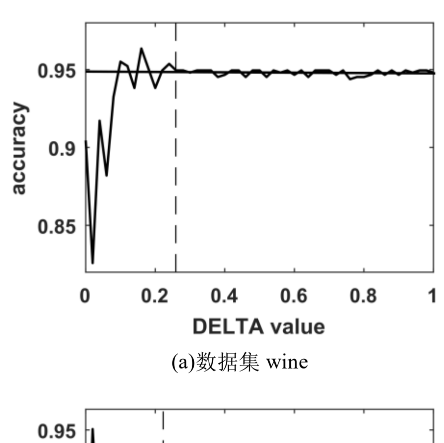
δ 的取值与聚类精度的关系如图 3 中(a)、(b)、(c)所示。

分析图 3 中(a)~(c)，横线代表原属性集的聚类精度，折现代表不同 δ 取值下对应属性约简的聚类精度，虚线代表饱和点。

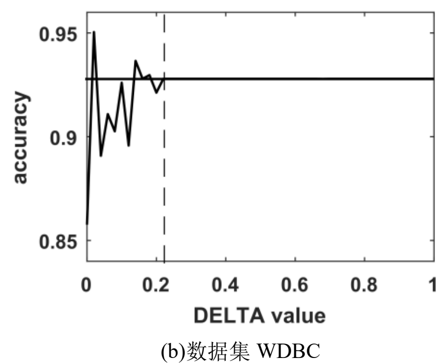
以饱和点为基准，将图 3 中(a)~(c)分为前后两部分，对同一数据集的原属性集而言，前半部分对应着在不同 δ 取值下的属性约简，后半部分对应着原属性集。

分析后半部分：在各图中，折线均在横线附近波动，这说明初选点的选择影响 K-means 算法聚类效果的稳定性。可以用折线相对于横线的波动程度表示 K-means 算法针对各数据集的聚类效果的稳定性。在(a)中，折线相对横线略有波动，这说明 K-means 算法对数据集 wine 的聚类效果较稳定；在(b)中，折线与横线完全契合，这说明 K-means 算法对数据集 WDBC 的聚类效果极稳定；在(c)中，折线相对横线波动较大，这是因为在本次实验的 δ 取值下，对数据集 sonar 得到的属性约简长度仍未趋于稳定。

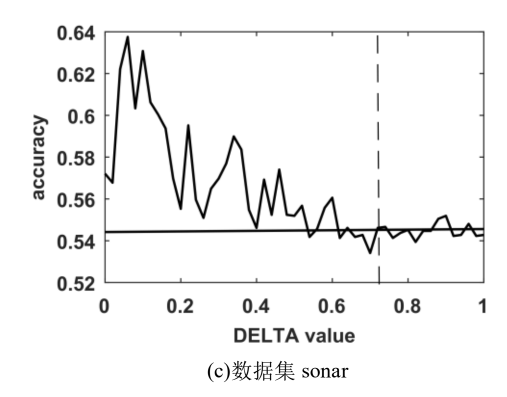
分析前半部分：首先，随着 δ 取值的增大，各图中的折线均存在高出横线后又低于横线的情况，这说明各数据集中均存在可以删除的冗余属性，且这些冗余属性会降低聚类精度。在(a)中，折线起点处于横线下方且随着 δ 的增大呈现递增状态，这说明在数据集 wine 中不存在仅靠个别属性就能将各样本进行正确聚类的情况，且冗余属性很少；在(b)中，在 δ 很小时，对应的折线部分远高于横线，然后迅速下降，这说明在数据集 WDBC 中存在仅靠个别属性就能将各样本进行正确归类的情况，冗余属性很多且严重影响聚类精度；在(c)中，折线大致处于横线上方且呈现明显的递减状态，这说明在数据集 sonar 中存在仅靠个别属性就能将各样本进行正确归类的情况，冗余属性很多且影响聚类精度。



(a)数据集 wine



(b)数据集 WDBC



(c)数据集 sonar

图 3 的取值与 K-means 聚类精度的关系

4.2.3 FARAIT 算法的实验结论

根据以上分析可知：在 δ 的取值合理的情况下，基于前向贪心的信息表属性约简算法能有效删除数据集中的冗余属性，在一定程度上消除冗余属性对聚类精度的干扰，优化聚类算法的性能。

考虑 K-Means 算法的不稳定性，相对来说，对于 K-means 算法而言，δ 在区间 $[0.14,0.18]$ 上取值时，K-means 算法的聚类效果较为理想。在对应的 δ 取值下，得到的属性约简个数小于原数据集且聚类精度高于原数据集。

4.3　FARAIT 算法与 PCA 算法的对比

根据 4.2 中的实验结论，对 FARAIT 算法设置 δ=0.16，对 PCA 算法设置阈值为 0.85，首先分别用这两种降维算法对数据集进行降维处理，最后用 K-means 算法对降维后的数据集进行聚类，统计各自对应的属性约简个数和聚类精度。其中，K-means 算法执行 20 次，聚类精度的最后结果取均值。实验结果如表 3 所示。

表 3　FARAIT 算法和 PCA 算法的实验结果对比

数据集		FARAIT 算法		PCA 算法	
名称	属性数	属性约简个数	聚类精度	属性约简个数	聚类精度
wine	13	9	0.9602	6	0.9302
WDBC	30	23	0.9283	10	0.9279
sonar	60	11	0.5962	15	0.5430
ionosphere	34	23	0.7123	15	0.7043
credit approval	14	14	0.7610	5	0.7142
german credit	24	16	0.5524	13	0.5552
WPBC	22	9	0.5758	7	0.5827
平均值	**28.14**	**15**	**0.7266**	**10.14**	**0.7082**

根据表 3 作出在两种降维算法下得到的属性约简个数和 K-means 算法的聚类精度的折线图，如图 4、5 所示。

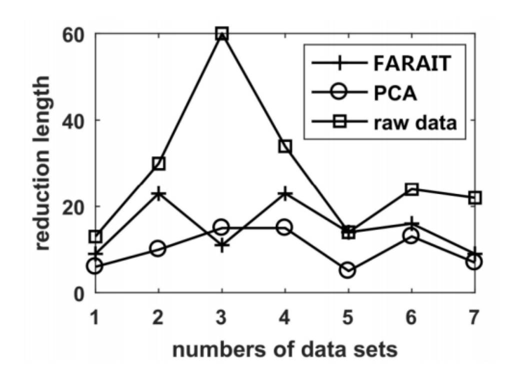


图 4　FARAIT 和 PCA 在属性约简个数上的对比

在图 4 中，相对而言，代表原始属性集的折线处于最上方，代表 FARAIT 算法的折线处于中间，代表 PCA 算法的折线处于最下方。这说明 FARAIT 算法和 PCA 算法都能有效地减少数据集的属性数，达到降维的目的。其次，相较 PCA 算法，FARAIT 算法得到的属性约简个数较多。

在图 5 中，相对而言，代表 FARAIT 算法的折线处于 PCA 算法的上方。这说明相较 PCA 算法，采用 FARAIT 算法降维后，K-means 算法的聚类精度较高。

综上可知：相较 PCA 算法，采用 FARAIT 算法降维后得到的属性约简个数较多，K-means 算法的聚类精度较高。

5　结束语

为了扩展邻域粗糙集对信息表的应用，设计了一种适用于聚类的无监督性信息表降维算法。在处理数值型信息表时，可采用该算法对数据进行预处理，用以删除信息表中的冗余信息，保持甚至提高聚类精度，优化聚类算法的性能。在本文的实验分析部分，可以看出 δ 的取值直接影响着 FARAIT 算法的效果，合适的 δ 的取值能让 FARAIT 算法的效果达到最好，不合适的 δ 的取值则会使 FARAIT 算法的效果一般，甚至很差。对于不同特性的聚类算法，δ 的取值为多少时 FARAIT 算法效果最好，从而让聚类效果较优，这个问题将在未来的工作中进行研究。

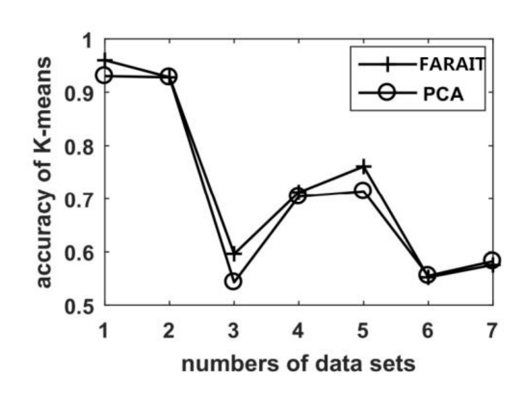


图 5　FARAIT 和 PCA 在 K-means 聚类精度上的对比

参考文献：

[1] 贺玲，蔡益朝，杨征. 高维数据聚类方法综述 [J]. 计算机应用研究, 2010, 27 (1): 23-26.

[2] Pawlak Z, So-Winski R. Rough set approach to multi-attribute decision analysis [J]. European Journal of Operational Research, 1994, 72 (3): 443-459.

[3] Zadeh LA. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic [J]. Fuzzy Sets & Systems, 1997, 90 (90): 111-127.

[4] Lin TY. Granular Computing on binary relations I: Data mining and neighborhood systems [J]. Rough Sets in Knowledge Discovery, 1998 (2): 165-166.

[5] Hu Q, Yu D, Liu J, Wu C. Neighborhood rough set based heterogeneous feature subset selection [J]. Information Sciences, 2008, 178 (18): 3577–3594.

[6] 王国胤. Rough 集理论与知识获取 [M]. 西安: 西安交通大学出版社, 2001: 147-156.

[7] 胡清华，于达人. 应用粗糙计算 [M]. 北京: 科学出版社, 2012.

[8] 刘遵仁，吴耿锋. 基于邻域粗糙模型的高维数据集快速约简算法 [J]. 计算机科学, 2012, 39 (10): 268-271.

[9] Liu Y, Huang W, Jiang Y, Zeng Z. Quick attribute reduct algorithm for neighborhood rough set model [J]. Information Sciences, 2014, 271 (7): 65-81.

[10] Chen H, Li T, Cai Y, et al. Parallel Attribute Reduction in Dominance-based Neighborhood Rough Set [J]. Information Sciences, 2016, 373: 351-368.

[11] Wang C, Shao M, He Q, et al. Feature subset selection based on fuzzy

neighborhood rough sets [J]. Knowledge-Based Systems, 2016, 111: 173-179.

[12] Chen Y, Zhang Z, Zheng J, et al. Gene selection for tumor classification using neighborhood rough sets and entropy measures. [J]. Journal of Biomedical Informatics, 2017, 67: 59-68.

[13] 王健, 冯健, 韩志艳. 基于流形学习的局部保持 PCA 算法在故障检测中的应用 [J]. 控制与决策, 2013 (5): 683-687.

[14] 刘丽敏, 樊晓平, 廖志芳, 等. 一种基于 L2, 1 范数的 PCA 维数约简算法 [J]. 计算机应用研究, 2013, 30 (1): 39-41.

[15] Hosoya H, Hyvärinen A. Learning Visual Spatial Pooling by Strong PCA Dimension Reduction [J]. Neural Computation, 2016, 28 (2): 1-16.